ISS Technology Update

Volume 8, Number 3

Keeping you informed of the latest ISS technology

HP Flex-10 and SR-IOV—What is the difference? 1

Three-phase power distribution in c7000 enclosures 3

PUE and DCE—Useful metrics for overall data center efficiency 6

Meet the Expert—Wayne Vuong 7

Recently published industry standard server technology papers 8

Contact us 8

HP Flex-10 and SR-IOV—What is the difference?

Improving I/O performonce

As virtual machine saftware enables higher efficiencies in CPU use, these same efficiency enablers place mare averhead an physical assets. HP Flex-10 Technalagy and Single Raat I/O Virtualizatian (SR-IOV) bath share the gaal af impraving I/O efficiency withaut increasing the averhead burden an CPUs and netwark hardware. Flex-10 and SR-IOV technalagies accamplish this gaal thraugh different appraaches. This article explares the differences in architecture and implementatian between Flex-10 and SR-IOV.

How these two technologies differ ond why it motters

HP Flex-10 Technology

Flex-10 technalagy is a hardware-based salutian that enables users ta partitian a 10 gigabit Ethernet (10GbE) cannectian and regulate the bandwidth af each partitian. HP Flex-10 is available anly with Virtual Cannect (VC), and is currently available far implementatian anly with supparted HP BladeSystem servers.

Flex-10 can be deplayed naw with HP BladeSystem servers, and with relatively little madificatian ta existing I/O architecture. Adding Flex-10-capable netwark devices and virtual cannect madules ta existing HP BladeSystem infrastructure allaws administratars ta take advantage af the cansiderable I/O cantral and refinement delivered by Flex-10 technalagy.

Using Flex-10 technalagy, administratars can canfigure a single BladeSystem 10Gb netwark part ta represent faur physical netwark interface cantrallers (NICs), alsa called FlexNICs, with a tatal bandwidth af 10Gbps. These faur FlexNICs appear ta the aperating system (OS) as discrete NICs, each with its awn driver. While the FlexNICs share the same physical part, traffic flaw far each ane is isalated with its awn MAC address and virtual lacal area netwark (VLAN) tags between the FlexNIC and

Downloaded from www.Manualslib.com manuals search engine

VC Flex-10 interconnect module. Using the VC interface, an administrator can set and control the transmit bandwidth available to each FlexNIC.

Each dual-port Flex-10 device supports up to eight FlexNICs, four on each physical port, and each Flex-10 Interconnect Module can support up to 64 FlexNICs.

Single Root I/O Virtualization (SR-IOV)

SR-IOV's ability to scale is a major advantage. The initial Flex-10 offering is based on the original PCIe definition that is limited to 8 PCI functions per given device (4 FlexNics per 10Gb port on a dual port device). With SR-IOV, there is a function called Alternative Route Identifiers (ARI) that allows expansion of up to 256 PCIe functions. The scalability inherent in the SR-IOV architecture has the potential to increase server consolidation and performance.

Another SR-IOV advantage is the prospect of performance gains achieved by removing the hypervisor from the main data path. Hypervisors are essentially another operating system providing CPU, memory, and I/O virtualization capabilities to accomplish resource management and data processing functions. The data processing functionality places the hypervisor squarely in the main data path.

In current I/O architecture, data communicated to and from a guest OS is routinely processed by the hypervisor. Both outgoing and incoming data is transformed into a format that can be understood by a physical device driver. The data destination is determined, and the appropriate send or receive buffers are posted. All of this processing requires a great deal of data copying, and the entire process creates serious performance overhead.

With SR-IOV, the hypervisor is no longer required to process, route, and buffer both outgoing and incoming packets. Instead, the SR-IOV architecture has exposed the underlying hardware to the guest OS, and while the guest OS still employs virtual Ethernet, it is now "paravirtualized." Paravirtualization lets the guest OS translate data by employing a virtual machine monitor (VMM). Calls to the hardware device drivers from the guest OS are replaced with calls to the VMM which handles the data processing tasks formerly performed by the hypervisor. This removes the hypervisor from the main data path and eliminates a great deal of performance overhead. The hypervisor continues to allocate resources and handle exception conditions, but it is no longer required to perform routine data processing.

Since SR-IOV is a hardware I/O implementation, it also uses hardware-based security and quality of service (QoS) features incorporated into the physical host server.

Summary

Advantages associated with using Flex-10 technology are significant. The ability to adjust transmit bandwidth by partitioning data flow makes 10GbE more cost effective and easier to manage. It is easier to aggregate multiple 1Gb data flows and fully utilize 10Gb bandwidth. The fact that Flex-10 is hardware based means that multiple FlexNICs are added without the additional processor overhead or latency associated with server virtualization (virtual machines). Significant infrastructure savings are also realized since additional server NIC mezzanine cards and associated interconnect modules may not be needed. See http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01608922/c01608922.pdf for more detailed information on Flex-10.

It is important to note that Flex-10 is an available HP technology for ProLiant BladeSystem servers, while SR-IOV is a released PCI-SIG specification at the beginning of the execution and adoption cycle. As such, SR-IOV cannot operate in current environments without significant changes to I/O infrastructure and the introduction of new management software. Once accomplished, these changes in infrastructure and software would let SR-IOV data handling operate in a much more native and direct manner, reducing processing overhead and enabling highly scalable PCI functionality.

For more information on the SR-IOV standard and industry support for the standard, go to the Peripheral Component Interconnect Special Interest Group (PCI-SIG) site at - http://www.pcisig.com.

Downloaded from www.Manualslib.com manuals search engine

For more information

For additional information, refer to the resources listed below.

Resource description	Web address
I/O Virtualization and Sharing	http://www.pcisig.com/developers/main/training_materials/get_document?doc_id=0cd682cb65b452d00e2047287b48369fc0d9bfa6
HP Virtual Connect Technology	http://h18004.www1.hp.com/products/blades/virtualconnect/index.html
HP Flex-10 Technology	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01608922/c01608922.pdf
Flex-10 VC Ethernet Module	http://h18004.www1.hp.com/products/blades/components/ethernet/10-10gb-f/index.html
HP NC532m Dual Part Flex-10 10GbE	http://h18004.www1.hp.com/products/servers/networking/nc532m/index.html?jumpid=reg_R1002_USEN
Implementing PCI I/O Virtualization Standards	http://download.microsoft.com/download/a/f/d/afdfd50d-6eb9-425e-84e1-b4085a80e34e/SYS-T311_WH07.pptx
10 Gigabit Ethernet technology for industry-standard servers, 2nd edition	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01608915/c01608915.pdf

Three-phase power distribution in c7000 enclosures

The HP BladeSystem c7000 Enclosure is available in models that use either 3-phase or single phase power to the enclosure. Enclosures that use 3-phase power can provide several advantages:

- More efficient power distribution. Delivering higher voltage 3-phase power to the enclosure results in less power loss during the transmission process.
- Elimination of PDUs. The AC Input Module of a 3-phase c7000 enclosure hooks directly to 3-phase power.
- Simpler wiring. Enclosures that use 3-phase power require only two cords connecting to the enclosure to support six modular power supplies rather than six cords coming from a PDU.

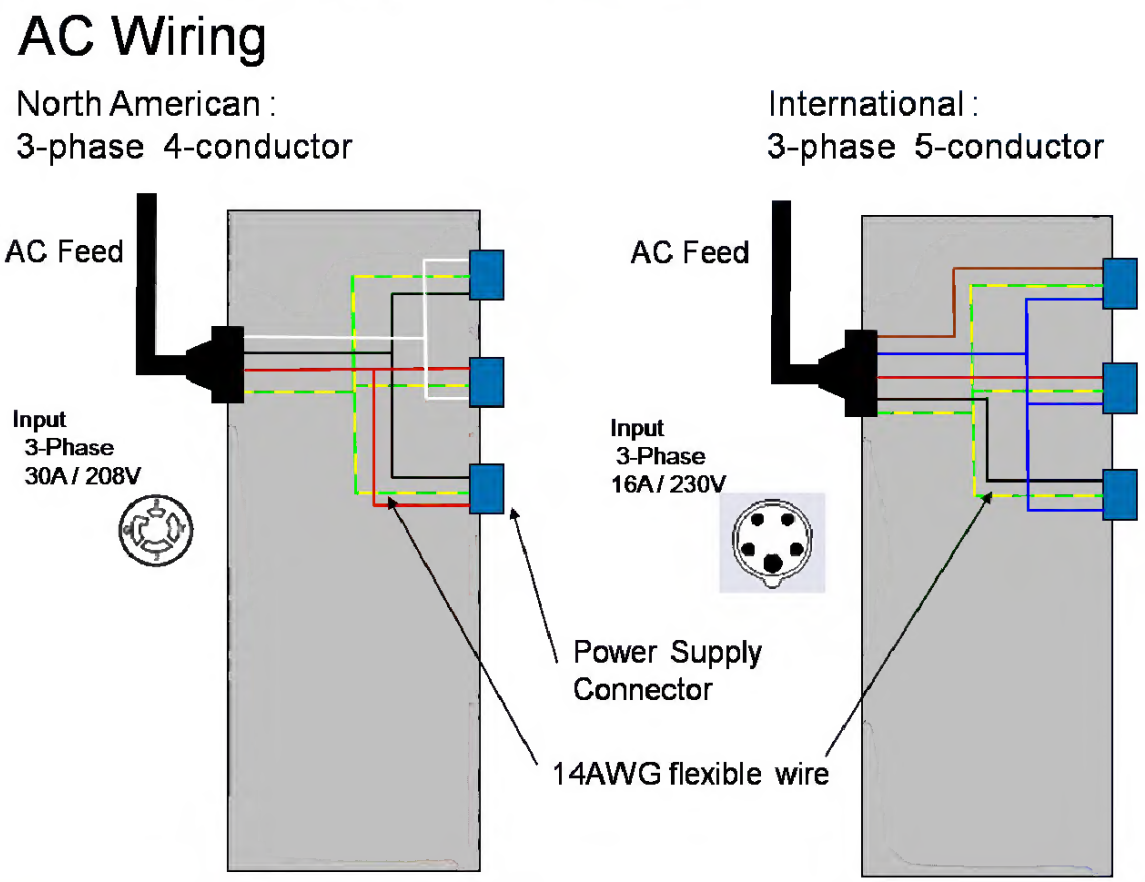
Distributing 3-phase power to the enclosure

There are two 3-phase feeds for the c7000 enclosure. They are marked on the input module as A and B, respectively. Feed A supplies power supplies 1, 2, and 3 while feed B supplies power supplies 4, 5, and 6.

As with all ProLiant servers that use AC power, the power supplies themselves use single phase current. Figure 2-1 illustrates that with the 3-phase c7000 enclosure, each power supply is connected to a single phase of the 208V (North America) or 230V (International) 3-phase feed.

Downloaded from www.Manualslib.com manuals search engine

Figure 2-1. Physically mapping 3-phase power to c7000 AC input module.



Colculoting 3-phose input power

Figure 2-2 shows how the power is electricolly mopped for both North Americo ond Internotionol 3-phose power into the single-phose power supplies.

Internotionol 3-phose power uses five conductors, including o true neutrol conductor, ollowing eoch of individuol phoses to deliver the full 230 volts. North Americon 3-phose power uses four conductors. The neutrol conductor for one phose is one of the line conductors for onother phose. Thus, only o differentiol voltoge ond not the full 208 volts is delivered to eoch of the power supplies. Additionolly, North Americon stondords stote thot circuits should not continuously drow more thon 80% of their roted current. Therefore, o 30 Amp circuit delivers only 24 Amps continuously. All of this must be token into considerotion when colculoting the totol input power from 3-phose circuits.

For o 30 Amp, 3-phose North Americon circuit, the input power delivered is colculoted os follows:

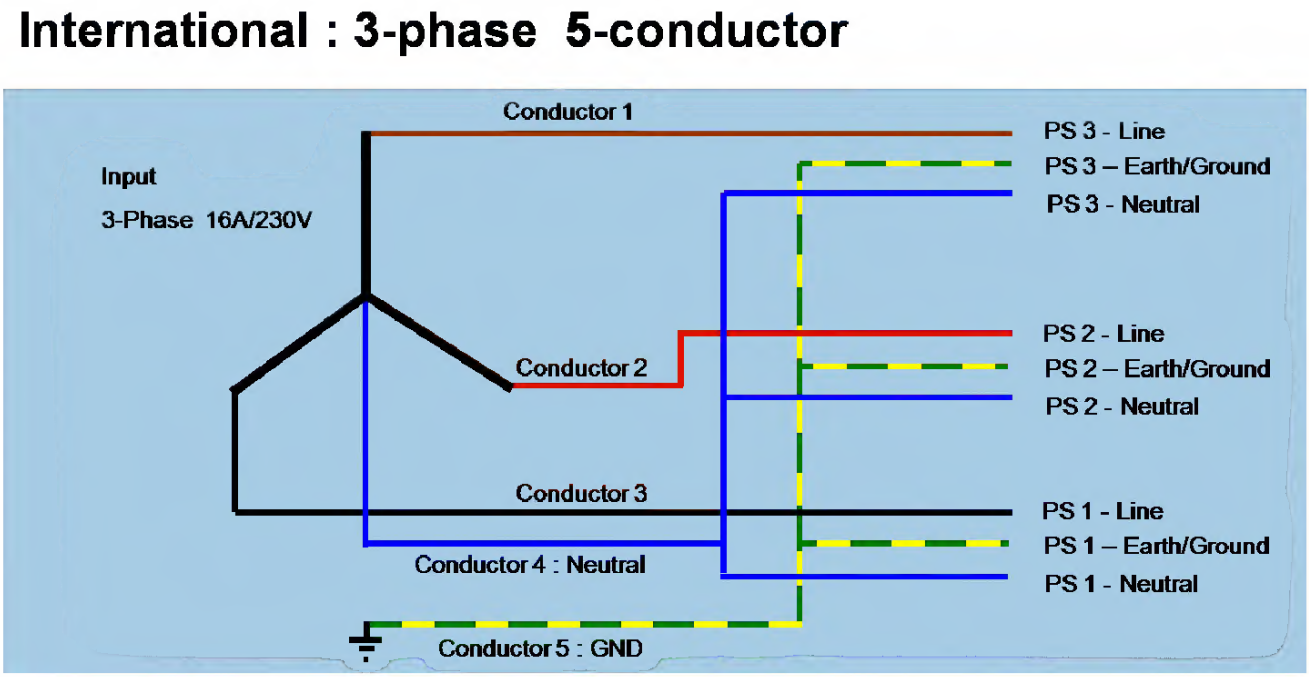
Continuous power = 1.732 x 208V x (.8 x 30A) = 8.6 kVA

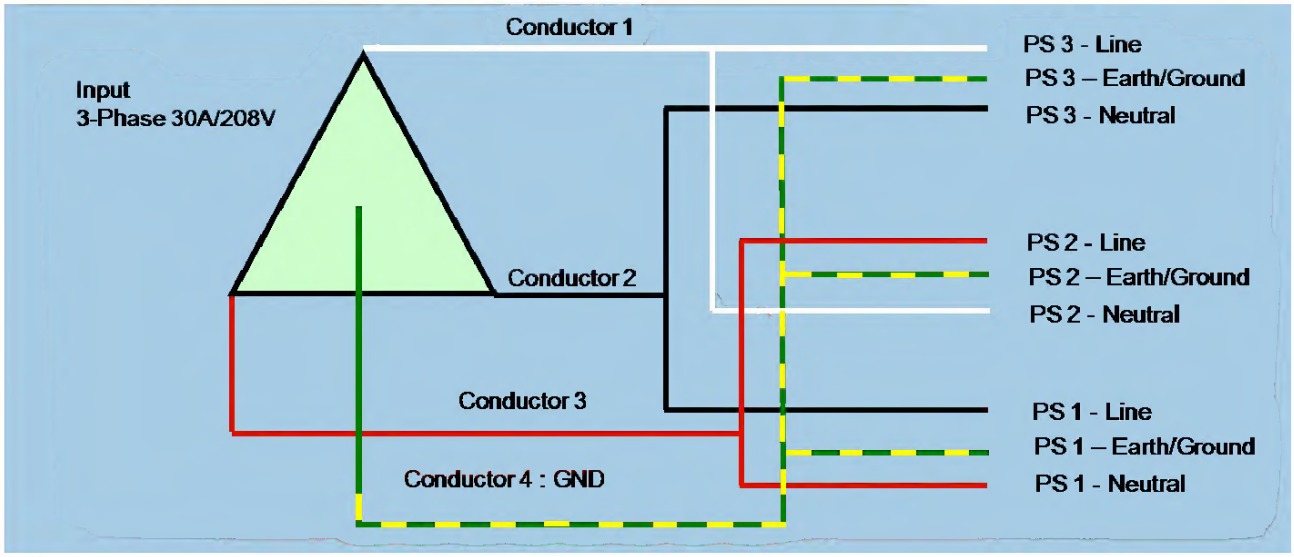
For o 16 Amp 3-phose Internotionol circuit, this would be the input power delivered:

Continuous power = 3 x 230V x 16A = 11.1 kVA

Downloaded from www.Manualslib.com manuals search engine

Figure 2-2. Electrical diagrams of 3-phase power to a c7000 AC input module





Additionol resources

For odditionol informotion on the topics discussed in this orticle, visit this website:

Resource	URL
HP BladeSystem c7000 Enclosure	http://h18004.www1.hp.com/products/blades/components/enclosures/c-class/c7000/

Downloaded from www.Manualslib.com manuals search engine

PUE and DCE—Useful metrics for overall data center efficiency

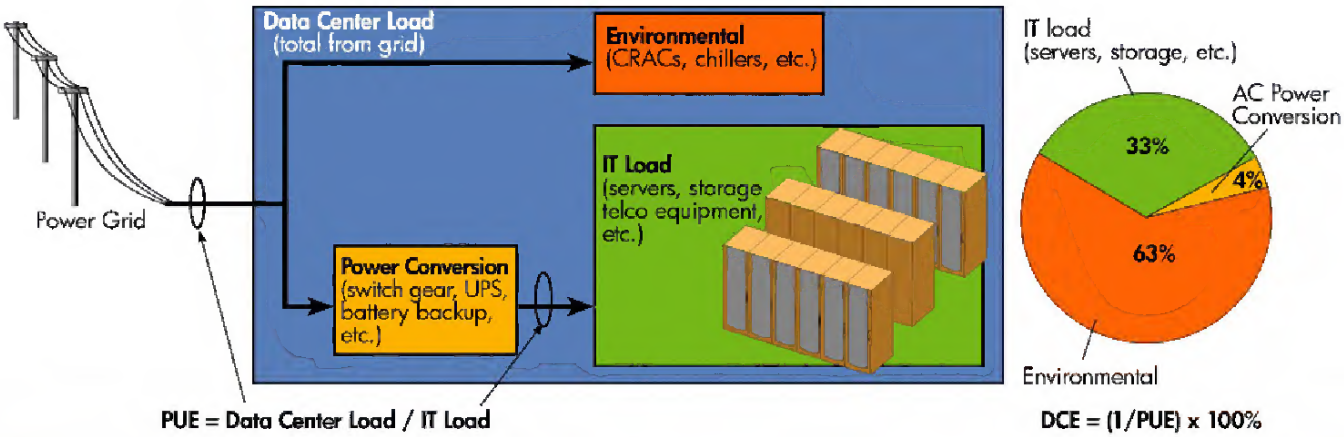
IT and facilities engineers can use two metrics to evaluate data center energy efficiency: Power Use Efficiency (PUE) and Data Center Efficiency (DCE). PUE and DCE are not intended to measure the efficiency of specific power and cooling systems within data center infrastructures. Rather, they are useful as "field" metrics to quantify overall data center efficiency.

As illustrated in Figure 3-1, PUE is the total power delivered to the facility divided by the net power that goes directly to the IT equipment. PUE can be used by an organization to benchmark its data center operations against the industry. It also offers a simple metric for determining compliance with governmental energy policies.

DCE—the inverse of PUE—is an alternative metric that is represented as a percentage. The DCE of a non-optimized data center may be 60%, while an optimized data center could have a DCE of 75% or higher.[1] Research by the Uptime Institute[2] indicates that 85% of data centers consume 2 kW for each kW used by IT equipment, with only a small portion going to power conditioning/conversion. This energy use results in a PUE of 3, and yields a DCE of 33 percent, meaning that only one-third of facility power is consumed by IT equipment (Figure 3-1 right).

Data centers with a PUE of 3 or greater typically have a grossly over-provisioned cooling system. Over-provisioning increases capital and recurring expenses and decreases utilization and efficiency, which results in a higher total cost of ownership (TCO). Consequently, alternate strategies are required to isolate root causes of inefficiency and provide remedial solutions. These strategies range from adopting the latest best practices for high-density computing platforms to creating a management layer that enables consolidation and an infrastructure that can dynamically provision power and cooling resources as needed.

Figure 3-1. PUE is the total data center load divided by the power that goes directly to IT equipment. DCE is the inverse percentage of PUE.



[1] Aebischer, B. Eubank, H., Tschudi, W., "Energy Efficiency Indicators for Data Centers," International Conference on Improving Energy Efficiency in Commercial Buildings "IEECB'04", 21 - 22 April 2004, Frankfurt (Germany). DCE is equivalent to CEE (Coefficient of Energy Efficiency).

[2] Belady, C., Malone, C., "Metrics to Characterize Data Center & IT Equipment Energy Use," Digital Power Forum, Richardson, TX (September, 2006).

Downloaded from www.Manualslib.com manuals search engine

Meet the Expert—Wayne Vuong

Wayne Vuang has been an Engineering Pragram Manager with Industry Standard Servers (ISS) far the last three years af his 12-year career at HP. Wayne has distinguished himself because af his hard wark and passian far getting custamer input ta engineering teams sa they can incarparate it inta their design decisians.

Engineering was a natural pragressian far Wayne because even at an early age he was very analytical, creative, and detail ariented. He enjays puzzles and math prablems. His favarite pastimes are fishing, billiards, card games, and playing the piana. He and his wife, Yisel, have a 4-year ald san, Janathan, wha is well pratected by their 12-year ald beagle, Maggie.

Name: Wayne Vuong
Title: Engineering Progrom Monoger, ISS
Years at HP: 12
University: BSEE, Rice University
MEE, Rice University
U.S. Patents: 2

A possion for customer input

Wayne is highly regarded far his suppart af the ISS Technical Advisary Cammittee (TAC), a prafessianal sacial netwark that facilitates interactian between HP custamers and ISS technalagists. Accarding ta his manager, Jeaff Krantz, "Wayne's invalvement with TAC yielded excellent custamer feedback that affected praduct decisians." In fact, Wayne says that almast all praduct design decisians are a result af custamer feedback received fram marketing, management, and direct custamer visits.

Helping drive the Common Design effort

Wayne alsa architected the successful Camman Design pracess that has been implemented in fifth-generatian (G5) and G6 PraLiant servers, and led the Camman Design team far G5 and midway inta G6. The Camman Design effart creates a server schematic that is directly shared by mast PraLiant design teams, significantly increasing design quality and efficiencies.

Synchranizing all platfarms with the same praven design virtually eliminates design errars that may accur if, far example, ane platfarm intentianally ar unintentianally implements a circuit differently. The Camman Design effart alsa prevents cammunicatian errars ar breakdawns where infarmatian may nat be disseminated praperly, which can result in a platfarm having a design quality issue.

Wayne says that getting the Camman Design pracess ta wark was very challenging because many believed that it cauld nat be dane. Alsa, variaus hurdles kept appearing that cauld have derailed the effart. Nevertheless, Wayne kept implementing warkaraunds ta keep the Camman Design visian alive.

In the end, having server platfarm teams use camman designs gave many engineers mare bandwidth ta facus an ather design activities. As these design activities were finished, they were alsa autamatically incarparated by the ather platfarm teams thraugh the Camman Design effart. This further reduced the design effart expended by each platfarm team, resulting in even mare free engineering bandwidth.

Impressed by ProLiont G6 servers' power efficiency

When asked which technalagies exemplify HP's leadership, Wayne remarked that the ability af PraLiant G6 servers ta dynamically balance pawer efficiency with perfarmance thraugh HP iLO (Onbaard Administratar) is phenamenal. Pawer and thermal cantrals embedded inta Onbaard Administratar reduce server pawer cansumptian by regulating pracessar speed and the aperatian af ather server campanents, such as fans, pawer supplies, and unpapulated I/O and memary slats.

Downloaded from www.Manualslib.com manuals search engine

Recently published industry standard server technology papers

Title	URL
Using InfiniBand for a Scalable Compute Infrastructure	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00593119/c00593119.pdf
Power Capping and Dynamic Power Capping for ProLiant Servers	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01549455/c01549455.pdf?jumpid=reg_R1002_USEN
Memory technology evolution: an overview of system memory technologies	http://h20000.www2.hp.cam/bc/dacs/suppart/SuppartManual/c00256987/c00256987.pdf
The Intel® processor roadmap for industry-standard servers, 9th Edition	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00164255/c00164255.pdf
10 Gigabit Ethernet technology for industry-standard servers, 2nd edition	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01608915/c01608915.pdf
Converged fabrics: Emerging technologies for simplifying data center infrastructure	http://h20000.www2.hp.cam/bc/dacs/suppart/SuppartManual/c01681871/c01681871.pdf
Data security in HP ProLiant servers using the Trusted Platform Module and Microsoft® Windows ® BitLocker™ Drive Encryption	http://h20000.www2.hp.cam/bc/dacs/suppart/SuppartManual/c01681891/c01681891.pdf
Electrical signal integrity considerations for HP BladeSystem	http://h20000.www2.hp.cam/bc/dacs/suppart/SuppartManual/c01712559/c01712559.pdf
Delivering an Adaptive Infrastructure with the HP BladeSystem c-Class architecture	http://h20000.www2.hp.cam/bc/dacs/suppart/SuppartManual/c01720656/c01720656.pdf

Industry standard server technical papers can be faund at www.hp.cam/servers/technalagy.

Contact us

Send camments abaut this newsletter ta TechCam@HP.cam.

Ta subscribe ta the ISS Technalagy Update newsletter, click mailta:techcam@hp.cam?subject=newsletter_subscriptian

Legal Notices© Copyright 2009 Hewlett-Packard Development Company, L.P. The information contoined herein is subject to chonge without notice. The only worronties for HP products and services are set forth in the express warronty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editoriol errors or omissions contained herein.

TC090411NL
April 2009

Downloaded from www.Manualslib.com manuals search engine